SkyStream Analytics

{Requirement Understanding document}

(By : Deepshikha Paty)

AGENDA:

- Understanding Business Requirements
- Understand Data
- Develop Solution Approach
- Implement and Validate

REPORT DESCRIPTION:

- The Global Airlines System manages a comprehensive dataset including air travel data such as airports, airlines, routes, and airplanes.
- To analyze the performance of airports, airlines, planes, and routes, various Key Performance Indicators (KPIs) are defined and implemented through a dashboard using visualization tools.

OBJECTIVES:

A company aims to utilize the Azure cloud platform to efficiently store and manage big data in various formats. The data will support business users with descriptive analytics using tools like Power BI. The data pipeline must ingest data from various sources into a structured data lake with zones for different stages of data processing. The solution must be scalable, cost-effective, and maintainable

BUSINESS REQUIREMENT:

- ✓ Migrate data from on-premise SQL server to cloud storage.
- ✓ Build a pipeline to transfer and process data.
- ✓ Connect Databricks to the storage account.
- ✓ Store aggregated data in the GOLD container of data lake storage for reporting with KPIs.
- ✓ Create dashboards using Power BI.

Key Performance Indicators (KPIs):

- The Below KPIs can include metrics such as total flights, passenger load factor, on-time performance, average delay, and route profitability.

1. **Total Flights**: Total number of flights.
2. **Passenger Load Factor**: Ratio of passengers carried to available seats.
3. **On-Time Performance**: Percentage of flights arriving on time.
4. **Average Delay**: Average delay time for flights.
5. **Route Profitability**: Profitability of specific routes.

UNDERSTANDING DATA:

Collected datasets in .csv files include:

1. **airports.csv**: Information about airports (Schema: Airport ID, Name, City, Country, IATA, ICAO, Latitude, Longitude, Altitude, Timezone, DST, Tz database time zone, Type, Source)
2. **airplanes.csv**: Information about airplanes (Schema: Name, IATA code, ICAO code)
3. **routes.csv**: Information about routes (Schema: Airline, Airline ID, Source airport, Source airport ID, Destination airport, Destination airport ID, Codeshare, Stops, Equipment)
4. **airlines.csv**: Information about airline companies (Schema: Airline ID, Name, Alias, IATA, ICAO, Callsign, Country, Active)

PROJECT ENVIRONMENT:

Microsoft Azure, commonly referred to as Azure, is a cloud computing service created by Microsoft for building, testing, deploying, and managing applications and services through Microsoft-managed data centers. Azure offers a wide range of cloud services, including computing, analytics, storage, and networking.

Data Factory: 

Data integration service that enables you to create, schedule, and manage data pipelines in Azure and beyond. It simplifies ETL (Extract Transform Load) scenarios like lifting-and-shifting first-generation information movement and transformation workloads at cloud destinations.

Data Lake Gen2: 

A data solution that combines the capabilities of a lake with the power of Azure Blob Storage, allowing you to store, serve, and analyze volumes of structured and unstructured information.

Databricks: 

Azure Databricks is a unified analytics platform built on top of Apache Spark. It helps with end-to-end data science solutions, including big data processing, machine learning experiments, and collaboration across various teams within an organization.

Delta Lake: 

A Delta Lake is an open-source storage layer designed to run on top of an existing data lake and improve its reliability, security, and performance. Delta Lakes support ACID transactions, scalable metadata, unified streaming, and batch data processing.

Logic Apps: 

Azure Logic Apps is a cloud-based service that allows you to automate workflows and integrate applications, data, and services across organizations. It provides a visual designer to create and manage workflows with pre-built connectors, enabling seamless communication between different systems without the need for extensive coding.

Power BI: 

Microsoft Power BI is a suite of business analytics tools designed to provide interactive visualizations and business intelligence capabilities with an interface simple enough for end-users to create their own reports and dashboards.

Security:

Microsoft Entra ID 

A cloud-based identity and access management service by Microsoft, providing secure single sign-on, multifactor authentication, and conditional access to safeguard user identities and access to applications.

Azure Key Vault: 

A cloud service for securely storing and accessing secrets, keys, and certificates, ensuring protection of sensitive information and simplifying key management processes.

IMPLEMENTATION STEPS:

- Go to SQL server and upload csv files for different tables.
- Add “who” columns (CreatedDate, ModifiedDate)
  - date datatype (default value is current date)
- Create storage account in Azure environment.
- Create Data factory.
- Azure setup integration run time -azure data factory
- Create a control table which has all table names
- Extract data from on-premise SQL server to azure ADLS storage -via ADF pipeline.

- Databricks Connect to ADLS Storage account-using service principal.
- Cleansing the data using databricks.
- Transformation on data in Databricks-using PySpark .
- Load the final aggregate data into ADLS storage (Delta table)
- Connect to Power BI to delta table for visualization, report generation.

PROJECT FLOW:

1. Data Ingestion

Azure Data Factory (ADF) Pipelines:

Define pipelines to pull data from sources (on-premise SQL server).

Ensure data lands in the Landing Zone (csv file format).

2. Landing Zone:

Data will be stored in csv file format.(as-it-is data)

Read the data through databricks and write in bronze zone in parquet format.

Remove the data from landing zone after writing in bronze zone.

3. Raw Data Storage (Bronze Zone)

Azure Data Lake Storage (ADLS):

Organize the raw data in a structured manner within the Bronze.

(partition the data by current date)

Maintain data in its original format for traceability and reprocessing if needed.

4. Data Processing and Transformation (Silver Zone)

Databricks:

Cleanse data to remove duplicates, handle missing values, and normalize data formats.

Enrich data by integrating additional relevant information.

Store the transformed data in the Silver Zone in a format optimized for querying

(in delta format and delta table)

5.Data Aggregation (Gold Zone)

Perform aggregations, complex transformations and calculations needed to support the KPIs.

Store the aggregated data in the Gold Zone in a format optimized for analytics.

(in delta format and delta table – use append mode)

6.Data Access and Visualization

Set up necessary views and stored procedures to facilitate data access.

Connect Power BI to these databases for visualization, dashboard and reports etc.

Conclusion:

By following this detailed project flow, the company can effectively leverage the Azure cloud platform to store and manage big data, enabling business users to perform descriptive analytics with tools like Power BI. The solution will be scalable, cost-effective, and easy to maintain, ensuring long-term success and data-driven decision-making.

THE END